



Home Flight



Tugas Besar MA4282 Kapita Selekta Statistika II

# Prediksi Delay Penerbangan Menggunakan Gaussian Process Regression

Search Destination

Find Flight

MATIAS JUDATAMA PARTAHI	10119027
SUSANTY SITOMPUL	10119060
FADHILAH SANA RACHMAPUTRI	10119085
YOSHUA ROBERTUS HARTONO	10119102
ANDIKA ZIDANE FATURRAHMAN	10119111



# Daftar Isi

- 01 Latar Belakang
- 02 Tujuan
- 03 *Gaussian Process Regression*
- 04 Fungsi Kernel
- 05 Optimisasi *Hyperparameter*
- 06 Distribusi Posterior
- 07 *Dynamic GPR*

- 08 *Confusion Matrix*
- 09 Metodologi Penelitian
- 10 Introduksi Data
- 11 Pembersihan Data
- 12 Analisis Data
- 13 Hasil Eksperimen
- 14 Kesimpulan

# Latar Belakang

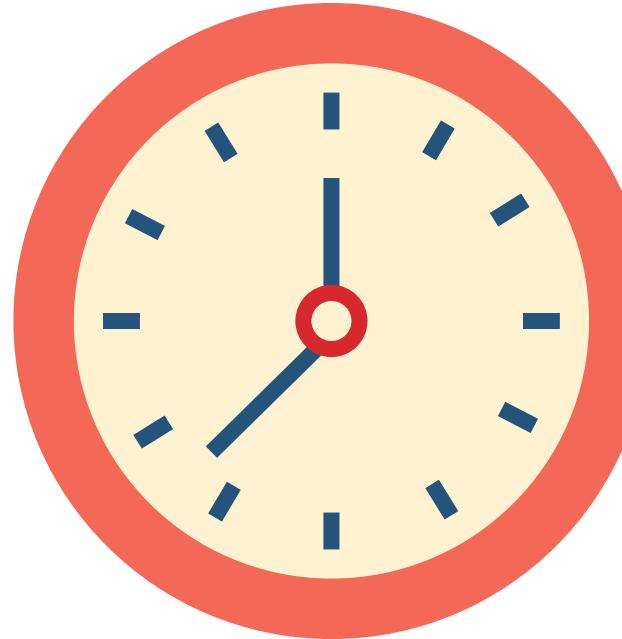


→  Departures

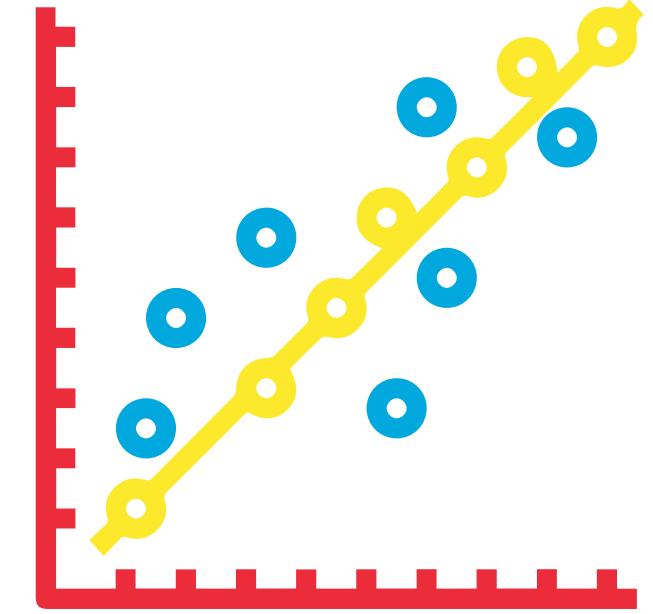
Berdasarkan data dari International Civil Aviation Organization (ICAO), terdapat lebih dari 100 ribu penerbangan komersial yang beroperasi setiap harinya di seluruh dunia pada tahun 2019.



Cuaca buruk serta kemungkinan terjadinya masalah teknis dapat menyebabkan keterlambatan pada penerbangan.



Dalam industri penerbangan, ketepatan waktu merupakan salah satu faktor yang berperan penting dalam menjamin kepuasan penumpang



Metode regresi *Gaussian Process* merupakan salah satu metode yang baik untuk mendapatkan hasil prediksi dari data yang besar



# Tujuan

1. Mendapatkan data prediksi delay keberangkatan pesawat untuk bulan April 2008 menggunakan regresi *Gaussian Process*.
2. Mendapatkan data prediksi delay keberangkatan pesawat dari salah satu maskapai penerbangan yaitu *United Airlines* untuk bulan April 2008 menggunakan regresi *Gaussian Process*.
3. Mendapatkan prediksi hari dengan delay paling minimum dengan menggunakan regresi *Gaussian Process*.
4. Mendapatkan data jumlah penerbangan yang termasuk kelompok *true positive*, *false positive*, *true negative*, dan *false negative* berdasarkan confusion matrix yang diperoleh.

# Gaussian Process Regression

Gaussian Process (GP) adalah proses stokastik sedemikian sehingga distribusi bersama dari setiap himpunan berhingga peubah acaknya adalah Gaussian multivariat. GP sepenuhnya ditentukan oleh fungsi rataan dan kovariansinya. Jika  $f(x)$  adalah proses pada input  $x$ , maka

$$f(x) \sim GP(m(x), k(x, x'))$$

dengan  $m(x) = E [f(x)]$  dan  $k(x, x') = E [(f(x) - m(x))(f(x') - m(x'))]$

Regresi dengan GP dilakukan melalui inferensi Bayesian untuk memperoleh distribusi posterior yang didasarkan pada fungsi prior (diasumsikan memiliki rataan 0) yang diperoleh dari data observasi. Kemudian, dengan input titik *test* baru, posterior dapat digunakan untuk memperoleh distribusi prediktif kondisional terhadap data observasi dan *test*.

# Fungsi Kernel

Salah satu bagian penting dalam GPR adalah pemilihan fungsi kernel dan estimasi parameter bebas  $\theta = \{\varphi_1, \varphi_2, \dots, \varphi_{n_\varphi}\}$  yang disebut dengan *hyperparameter*. Selain itu, nilai fungsi kernel  $k(x_i, x_j)$  yang menyatakan kovariansi antara dua elemen  $(x_i, x_j)$  memiliki syarat perlu dan cukup yaitu matriks kovariansinya semidefinit positif untuk semua kemungkinan *input* agar dianggap sebagai fungsi kernel yang valid.

Dalam eksperimen ini, digunakan kernel squared exponential ARD dengan persamaan berikut:

$$k(x, x') = \varphi_1^2 \exp(-(x - x')^T P^{-1}(x - x'))$$

dengan  $P = \text{diag}(\varphi_2^2, \dots, \varphi_{n_\varphi}^2)$

# Optimisasi Hyperparameter

Salah satu cara untuk menentukan *hyperparameter* yang paling sesuai pada model GP adalah dengan memaksimumkan fungsi *log marginal likelihood* dari data yang diobservasi. Fungsi *log marginal likelihood* adalah sebagai berikut:

$$\mathcal{L}(\theta) = \log(p(y)) = -\frac{1}{2}\mathbf{y}^T K_n^{-1} \mathbf{y} - \frac{1}{2}\log(|K_n|) - \frac{n}{2}\log(2\pi)$$

dengan  $\theta$  himpunan dari *hyperparameter* yang akan dioptimisasi dan  $K_n$  menyatakan kernel. Untuk memperoleh  $\theta$  yang memaksimumkan fungsi *log marginal likelihood*, digunakan metode *Gradient Descent*. Dengan  $\theta_0$  sebagai tebakan awal, maka  $\theta$  dapat diperbaharui dengan persamaan:

$$\theta_n = \theta_{n-1} - \eta \nabla \mathcal{L}(\theta_{n-1})$$

Pembaharuan  $\theta$  dilakukan hingga memenuhi kondisi:

$$\|\theta_n - \theta_{n-1}\| \leq \zeta, 0 < \zeta \leq 1$$

# Distribusi Posterior

Misalkan terdapat data observasi  $\{(x_{o,i}, y_{o,i})\}_{i=1}^n$ . Untuk memperoleh distribusi prediktif posterior dari  $f$  di titik baru  $x_*$ , digunakan inferensi Bayes sehingga diperoleh

$$p(f(x_*) \mid \{(x_{o,i}, y_{o,i})\}_{i=1}^n) \sim N(\mu_*, \sigma_*^2)$$

dengan

$$\mu_* = k(x_*, \mathbf{x}_o)(K_{o,o} + \sigma_o^2 I)^{-1} \mathbf{y}_o$$

$$\sigma_*^2 = k(x_*, x_*) - k(x_*, \mathbf{x}_o)(K_{o,o} + \sigma_o^2 I)^{-1} k(\mathbf{x}_o, x_*)$$

$$K_{o,o} = \mathbf{K} = \begin{pmatrix} k(\mathbf{x}_{o,1}, \mathbf{x}_{o,1}) & \dots & k(\mathbf{x}_{o,1}, \mathbf{x}_{o,n}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_{o,n}, \mathbf{x}_{o,1}) & \dots & k(\mathbf{x}_{o,n}, \mathbf{x}_{o,n}) \end{pmatrix}$$

Komponen  $\{(\mathbf{K} + \sigma_o^2 I)^{-1} \mathbf{y}_o\}$  pada fungsi rataan dari distribusi prediktif  $\mu(x)$  adalah konstanta sehingga dapat ditulis sebagai kombinasi linear dari kernel yaitu

$$\mu(x) = \sum_{i=1}^n w_i k(x_*, x_{o,i}) \text{ dengan } w_i = ((\mathbf{K} + \sigma_o^2 I)^{-1} \mathbf{y}_o)_i \text{ adalah bobot dan } \{(\mathbf{K} + \sigma_o^2 I)^{-1} \mathbf{y}_o\}$$

adalah fungsi basis

# Dynamic GPR

Dalam *Dynamic GPR*, yang menjadi *input* adalah data  $\{(x_{o,i}, y_{o,i})\}_{i=1}^N$ , tebakan awal  $\theta_0$ , pusat kelompok hasil *kmeans* (metode mengklasker data untuk memperoleh basis pada data observasi; dilakukan karena jumlah data yang besar):  $\{c_{b,i}\}_{i=1}^M$  dengan  $M \ll N$ , dan  $n = 0, m_0 = 0, S_0 = k(\mathbf{c}_b, \mathbf{c}_b; \theta_0)$

Algoritma:

1. Perbaharui  $\mathbf{m}_{n+1}, \mathbf{S}_{n+1}$  dengan data observasi ke  $n+1$  yaitu  $(x_{o,n+1}, y_{o,n+1})$
2. Perbaharui  $\theta_{n+1}$
3. Prediksi nilai  $f$  di lokasi baru  $x_*$
4. Bila  $n+1 < N$ , tetapkan  $n = n+1$  dan kembali ke langkah 1. Selain itu, selesai

*Output* yang didapat dari algoritma adalah  $m_N, S_N, \theta_N$

# Confusion Matrix

*Confusion matrix* adalah sebuah matriks yang dapat digunakan untuk mengevaluasi performa dari sebuah model machine learning dalam klasifikasi.

Keempat elemen pada confusion matrix adalah sebagai berikut:

1. **True Positive (TP)** yang menyatakan jumlah prediksi positif yang benar.
2. **False Positive (FP)** yang menyatakan jumlah prediksi positif yang salah (hasil prediksi positif, aktual negatif).
3. **True Negative (TN)** yang menyatakan jumlah prediksi negatif yang benar.
4. **False Negative (FN)** yang menyatakan jumlah prediksi negatif yang salah (hasil prediksi negatif, aktual positif).

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive Type 1 Error
	Negative	False Negative Type 2 Error	True Negative

# Metodologi Penelitian

Pada tugas besar ini digunakan kernel ARD dengan parameter sebagai berikut:

10

Sigma Awal

0.2

Bobot  
setiap Fitur

10 &  
25

Jumlah  
Cluster

0.01

*Learning  
Rate*

32

*Mini Batch*  
(dengan pengambilan  
sampel secara random)

Untuk memvalidasi model digunakan metrik Mean Squared Error (MSE) untuk prediksi waktu keterlambatan dan akurasi untuk prediksi klasifikasi suatu penerbangan terlambat atau tidak. Apabila hasil prediksi waktu keterlambatan kurang dari 15 menit, maka akan dikategorikan menjadi 'tidak terlambat'. Namun, jika sebaliknya, maka akan dikategorikan menjadi terlambat.

# Introduksi Data

Pada eksperimen ini, digunakan data penerbangan tiap maskapai pada tahun 2008. Data terdiri dari 2389217 baris dengan variabel:

1. *Year*: Tahun 2008
2. *Month/Bulan*
3. *DayofMonth*: Hari dalam bulan (1-31)
4. *DayofWeek*: Hari dalam minggu (1-7)  
dengan hari ke - 1 adalah Senin dan hari  
ke - 7 adalah Minggu
5. *DepTime*: Waktu keberangkatan aktual  
pesawat dalam format jam menit
6. *CRSDepTime*: Jadwal keberangkatan  
pesawat dalam format jam menit
7. *ArrTime*: Waktu mendarat aktual  
pesawat dalam format jam menit
8. *CRSArrTime*: Jadwal mendarat pesawat  
dalam format jam menit
9. *UniqueCarrier*: Kode unik maskapai
10. *FlightNum*: Nomor penerbangan
11. *TailNum*: Nomor ekor pesawat
12. *ActualElapsedTime*: Waktu penerbangan  
dihitung dari *ArrTime* – *DepTime*
13. *CRSElapsedTime*: Waktu penerbangan  
dihitung dari *CRSArrTime* – *CRSDepTime*
14. *AirTime*: Waktu penerbangan
15. *ArrDelay*: Waktu keterlambatan  
keberangkatan pesawat, dihitung dari  
*ArrTime* – *CRSArrTime*
16. *DepDelay*: Waktu keterlambatan  
mendaratnya pesawat, dihitung dari  
*DepTime* – *CRSDepTime*

# Introduksi Data

Pada eksperimen ini, digunakan data penerbangan tiap maskapai pada tahun 2008. Data terdiri dari 2389217 baris dengan variabel:

- 17. *Origin*: Asal bandara (kode IATA)
- 18. *Dest*: Tujuan bandara (kode IATA)
- 19. *Distance*: Jarak antara *Origin* dan *Dest* (mil)
- 20. *TaxiIn*: Interval waktu antara pesawat mendarat dengan parkir di terminal (menit)
- 21. *TaxiOut*: Interval waktu antara pesawat keluar dari terminal dengan keberangkatan pesawat di landasan (menit)
- 22. *Cancelled*: Variabel biner yang menyatakan apakah penerbangan dibatalkan

- 23. *CancellationCode*: Alasan penerbangan dibatalkan (A : maskapai, B: cuaca, C: NAS, D: keamanan)
- 24. *Diverted*: Variabel biner yang menyatakan apakah penerbangan mengalami perubahan rute
- 25. *CarrierDelay*: Waktu keterlambatan karena maskapai (menit)
- 26. *WeatherDelay*: Waktu keterlambatan karena cuaca (menit)
- 27. *NASDelay*: Waktu keterlambatan karena NAS (menit)
- 28. *SecurityDelay*: Waktu keterlambatan karena pengawasan (menit)
- 29. *LateAircraftDelay*: Waktu keterlambatan karena pesawat (menit)

# Pembersihan Data

Pada data penerbangan, terdapat nilai kosong (NaN) di beberapa fitur:



Fitur	Jumlah Nilai Kosong	Presentase Nilai Kosong (%)
<i>CancellationCode</i>	2324775	97.3028
<i>CarrierDelay</i>		
<i>WeatherDelay</i>		
<i>NASDelay</i>	1804634	75.53244
<i>SecurityDelay</i>		
<i>LateAircraftDelay</i>		
<i>ArrTime</i>		
<i>ActualElapsedTime</i>	70096	2.933848
<i>AirTime</i>		

# Pembersihan Data

Pada data penerbangan, terdapat nilai kosong (NaN) di beberapa fitur:

Fitur	Jumlah Nilai Kosong	Presentase Nilai Kosong (%)
<i>ArrDelay</i>	70096	2.933848
<i>TaxiIn</i>		
<i>DepTime</i>		
<i>DepDelay</i>	64442	2.697202
<i>TaxiOut</i>		
<i>TailNum</i>	42542	1.776816
<i>CRSElapsedTime</i>	407	0.017035

Dari tabel diatas, terdapat 16 variabel yang memiliki nilai kosong dengan persentase paling banyak pada fitur *CancellationCode*



# Pembersihan Data

- 64730 observasi dibuang dari fitur *DepDelay* dan *CRSElapsedTime* karena bernilai kosong.
- Nilai kosong pada fitur *CarrierDelay*, *WeatherDelay*, *NASDelay*, *SecurityDelay*, dan *LateAircraftDelay* diganti dengan nilai 0, karena nilai kosong yang berada pada fitur-fitur tersebut disebabkan salah satu fitur di antaranya telah diisi suatu nilai (dalam tiap observasi terjadi suatu delay yang disebabkan oleh setidaknya salah satu dari lima fitur tersebut).
- Diasumsikan bahwa variabel berikut tidak diketahui,
  - Waktu aktual keberangkatan,
  - Waktu aktual mendarat,
  - Waktu *taxis*,
  - Penerbangan dari destinasi ke tujuan, serta
  - Keterlambatan saat mendarat
- Fitur terkait yaitu *ArrTime*, *DepTime*, *ActualElapsedTime*, *AirTime*, *ArrDelay*, *TaxiIn*, dan *TaxiOut* tidak digunakan dalam pemodelan.
- Variabel *Year*, *Cancelled*, dan *CancellationCode* tidak digunakan karena variansi dari variabel tersebut bernilai 0

# Pembersihan Data

Variabel kategori *Origin* dan *Dest* akan dipetakan menjadi bilangan dengan menggunakan pemetaan *M-probability estimate of likelihood* (Micci-Barreca, 2001) berikut:

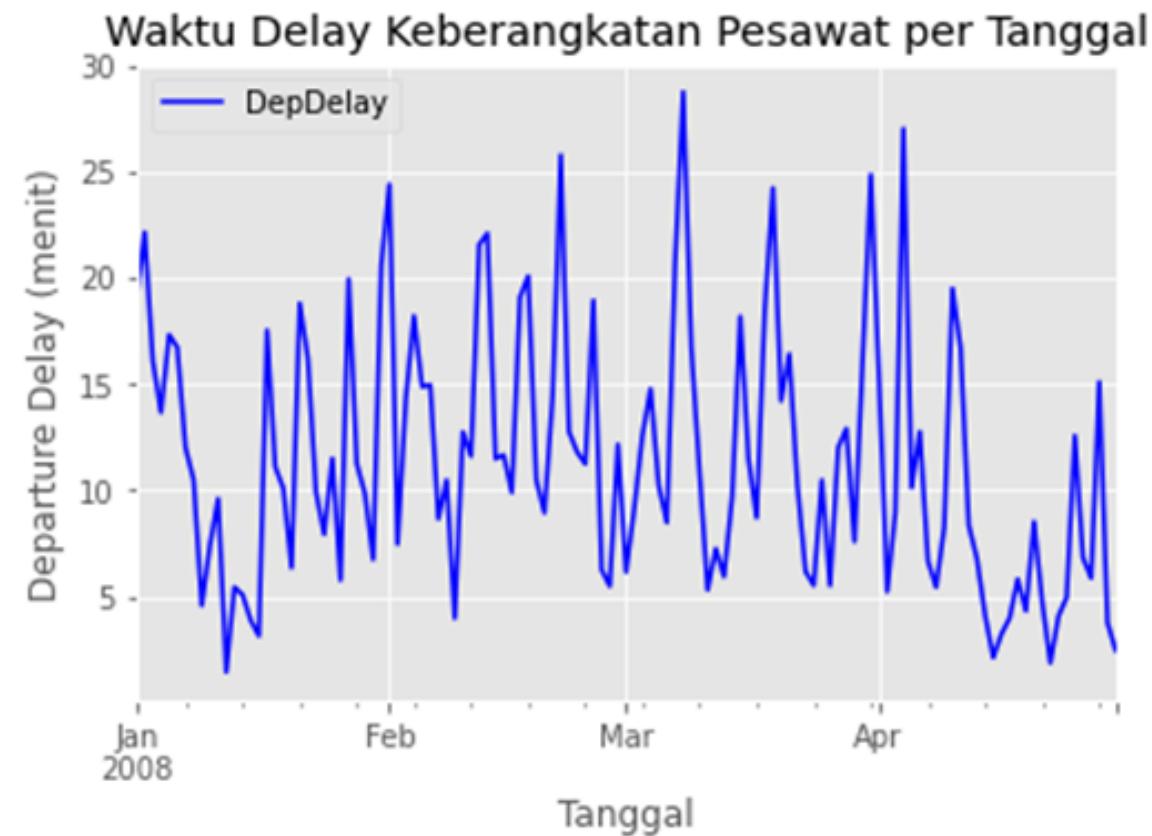
$$f(cat_1) = wP(y|cat_1) + (1 - w)P(y)$$

$$w = \frac{n}{n + m}$$

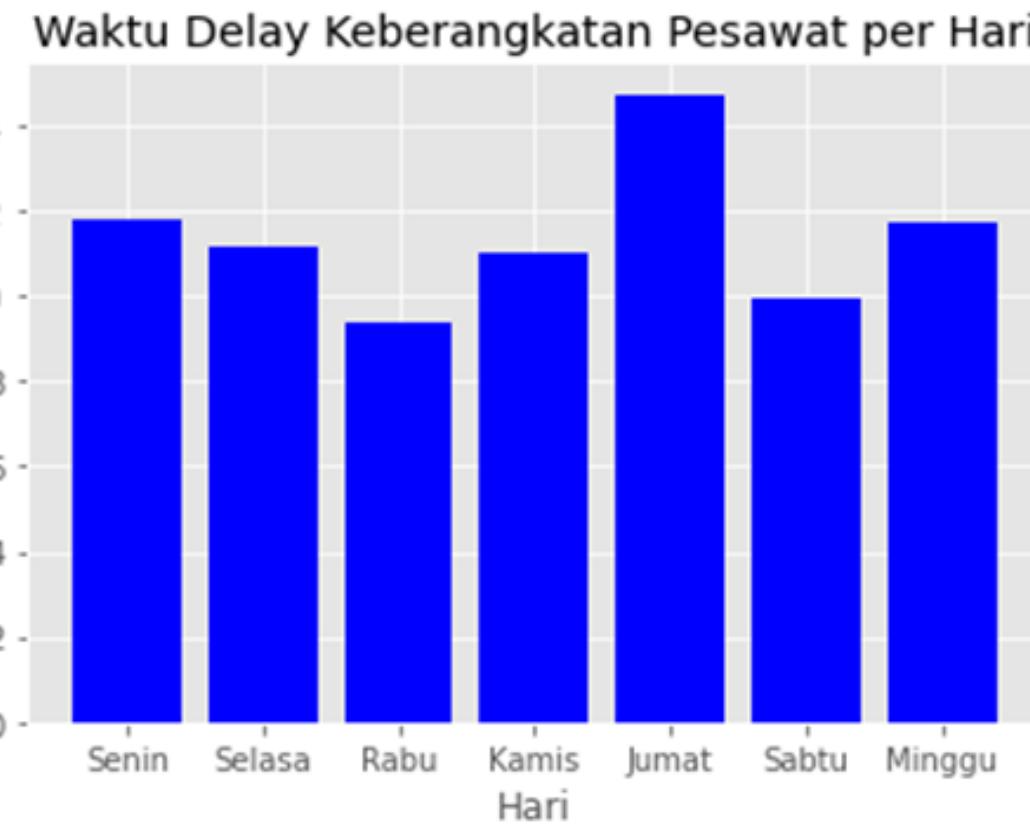
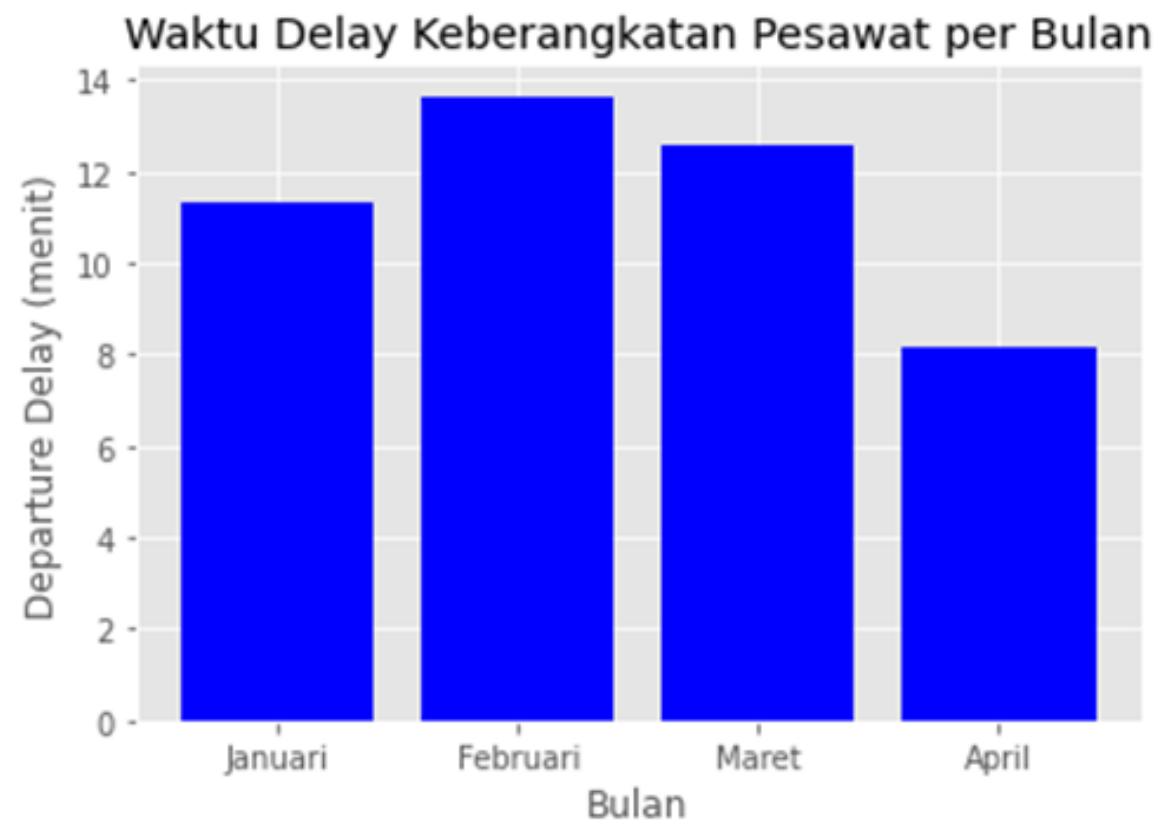
dengan,

- $cat_1$ , nilai kategori,
- $w$ , bobot penghalusan,
- $n$ , jumlah observasi yang mengandung kategori  $cat_1$
- $m$ , faktor smooth,
- $P(y | cat_1)$ , rerata DepDelay diberikan nilai kategori  $cat_1$
- $P(y)$ , rerata DepDelay secara keseluruhan.

# Analisis Data



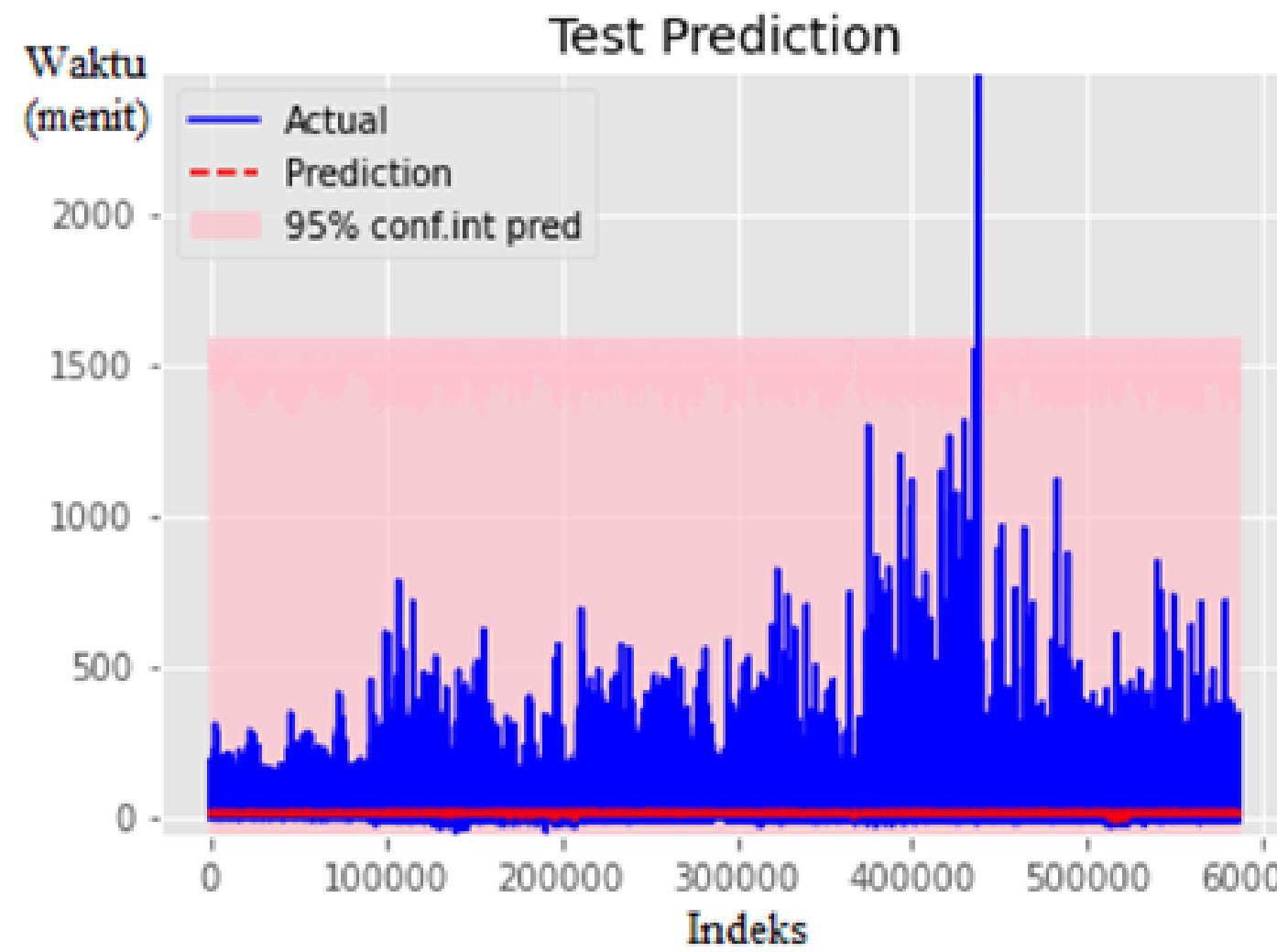
Pada bulan Maret 2008, rerata waktu keberangkatan pesawat mengalami keterlambatan selama hampir 30 menit.



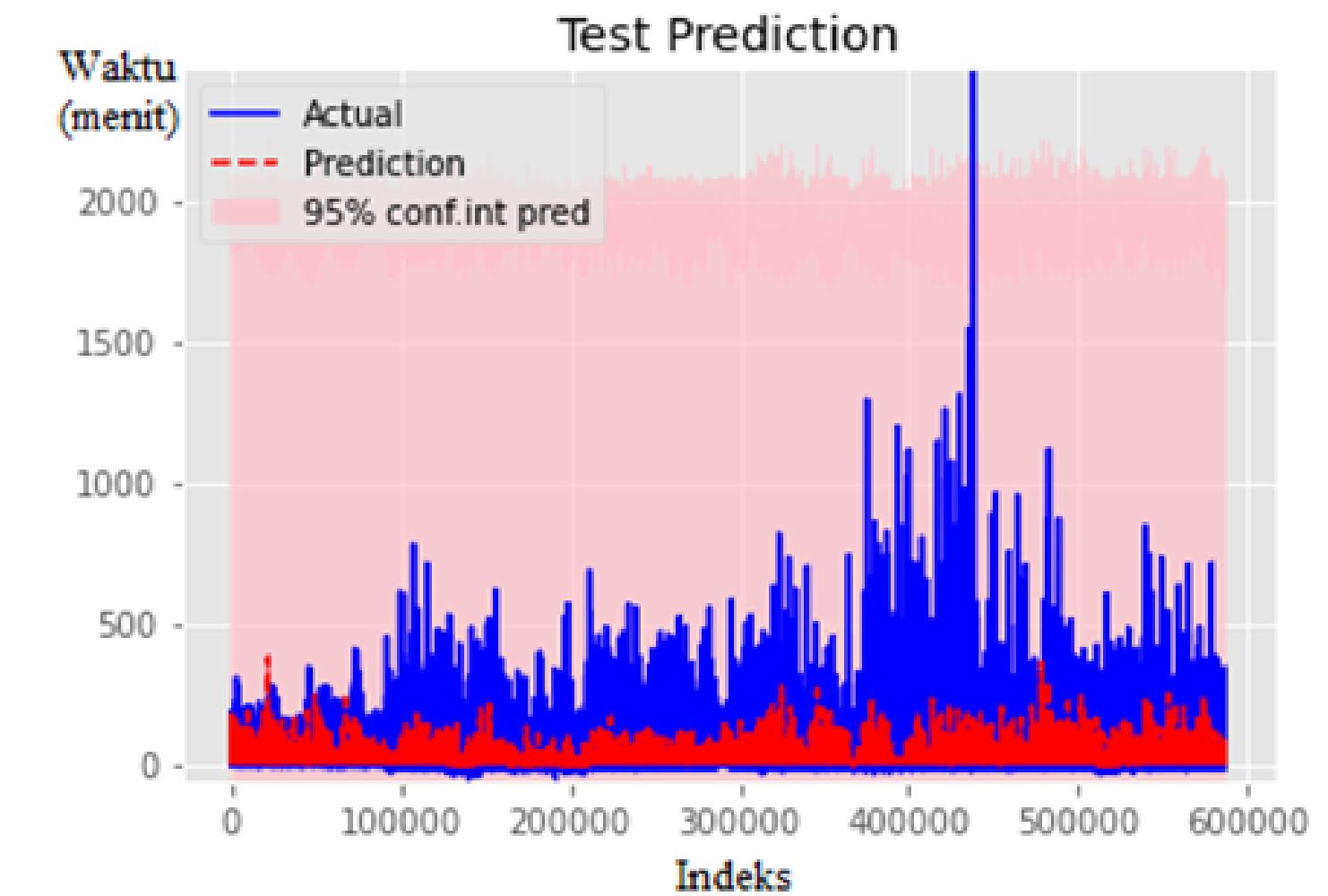
Waktu delay keberangkatan pesawat tertinggi tercapai pada bulan Februari dan hari Jumat. Lalu, terdapat tren menurun dari bulan Februari ke April serta tren menaik dari hari Rabu – Jumat dan Sabtu – Minggu.

# Hasil Eksperimen

Data dibagi menjadi 2 yakni data training yang berisi 1.732.398 observasi dari bulan Januari hingga Maret dan data test yang berisi 586.723 observasi pada bulan April.



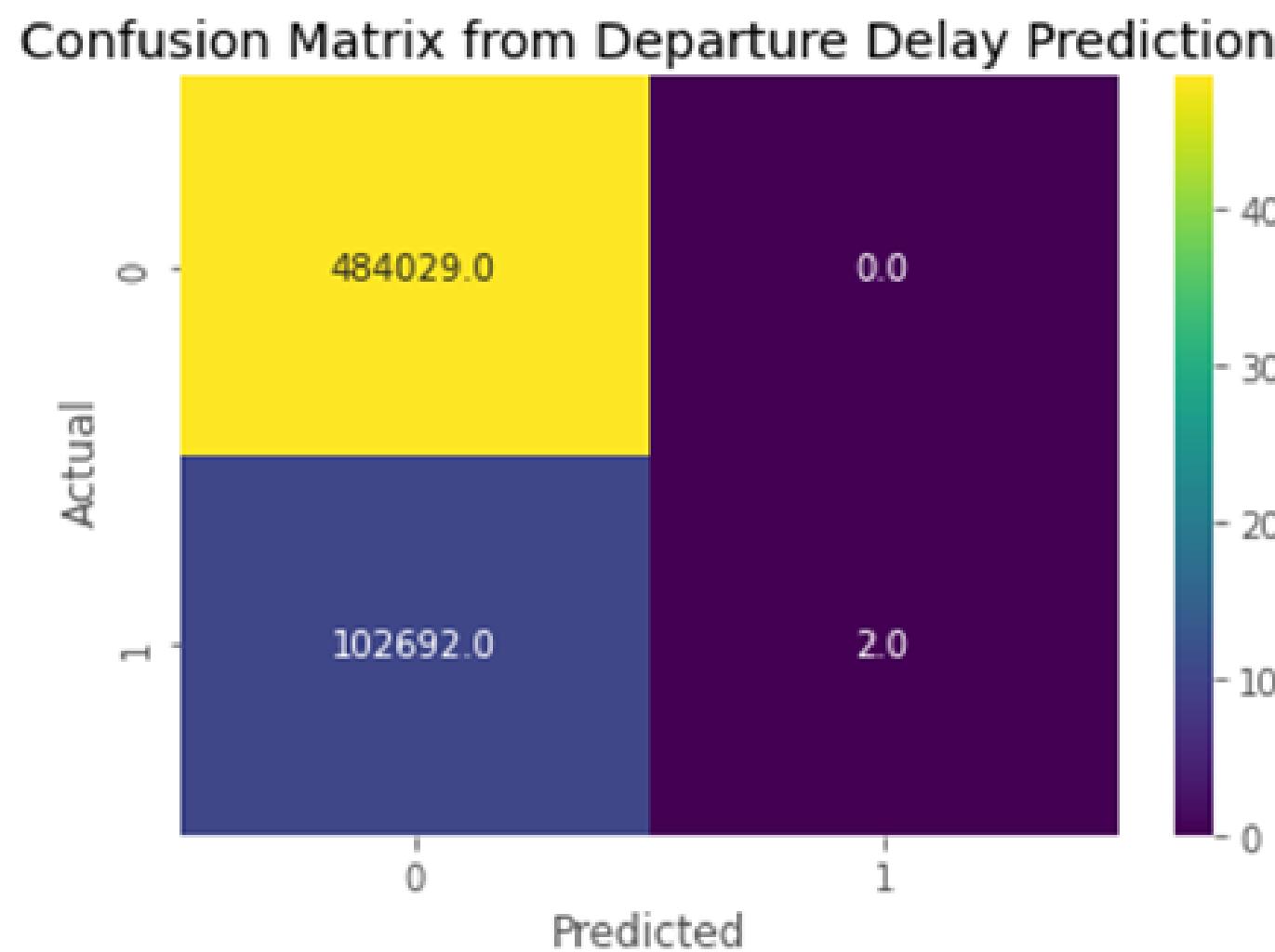
Jumlah Titik Basis 10



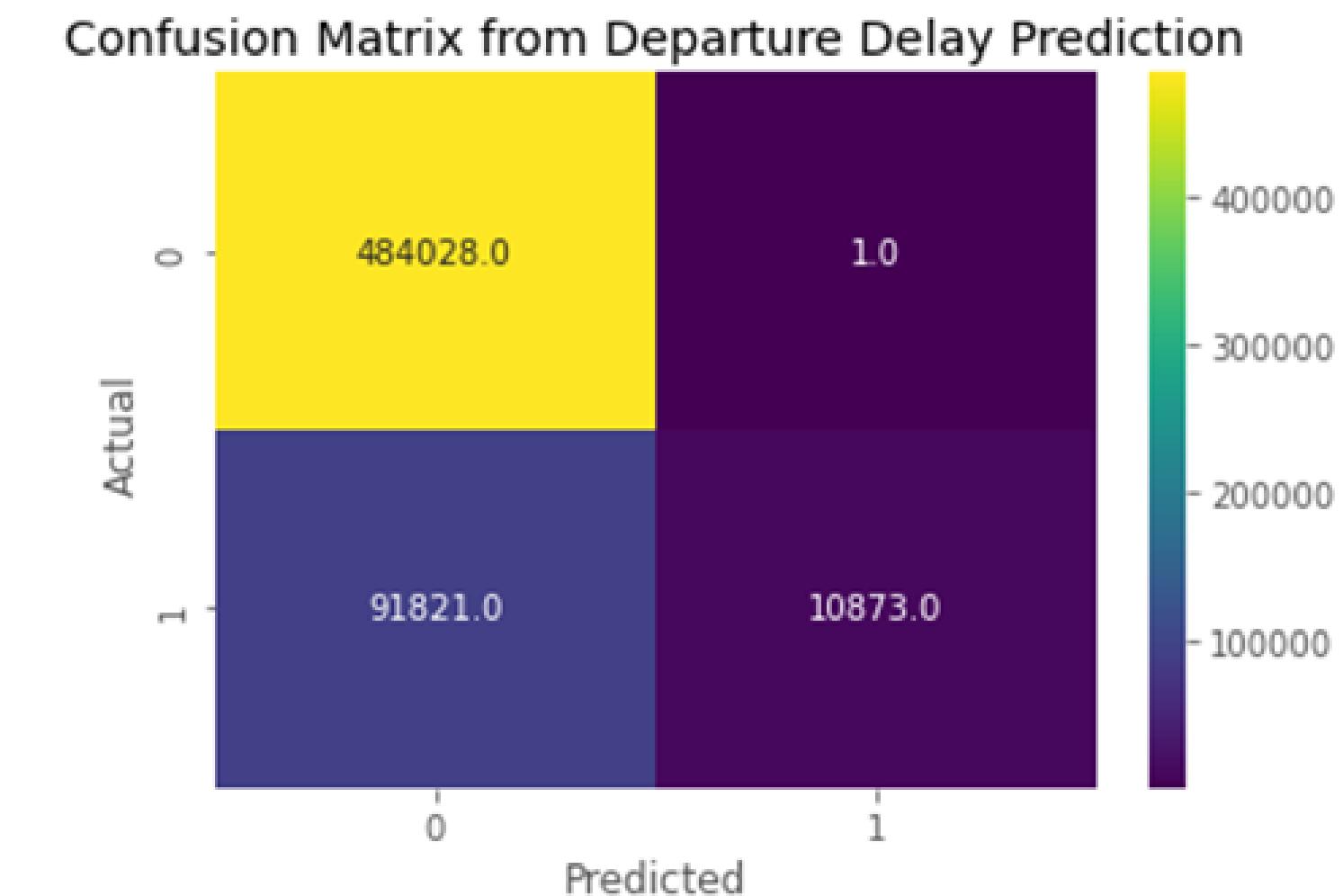
Jumlah Titik Basis 25

# Hasil Eksperimen

Pada *confusion matrix*, model dengan titik basis 10 tidak menghasilkan *false positive* namun hanya memprediksi 2 penerbangan dengan keberangkatan yang terlambat. Sementara, model dengan titik basis 25 hanya menghasilkan 1 *false positive* dan mampu memprediksi 10.873 penerbangan dengan keberangkatan yang terlambat.



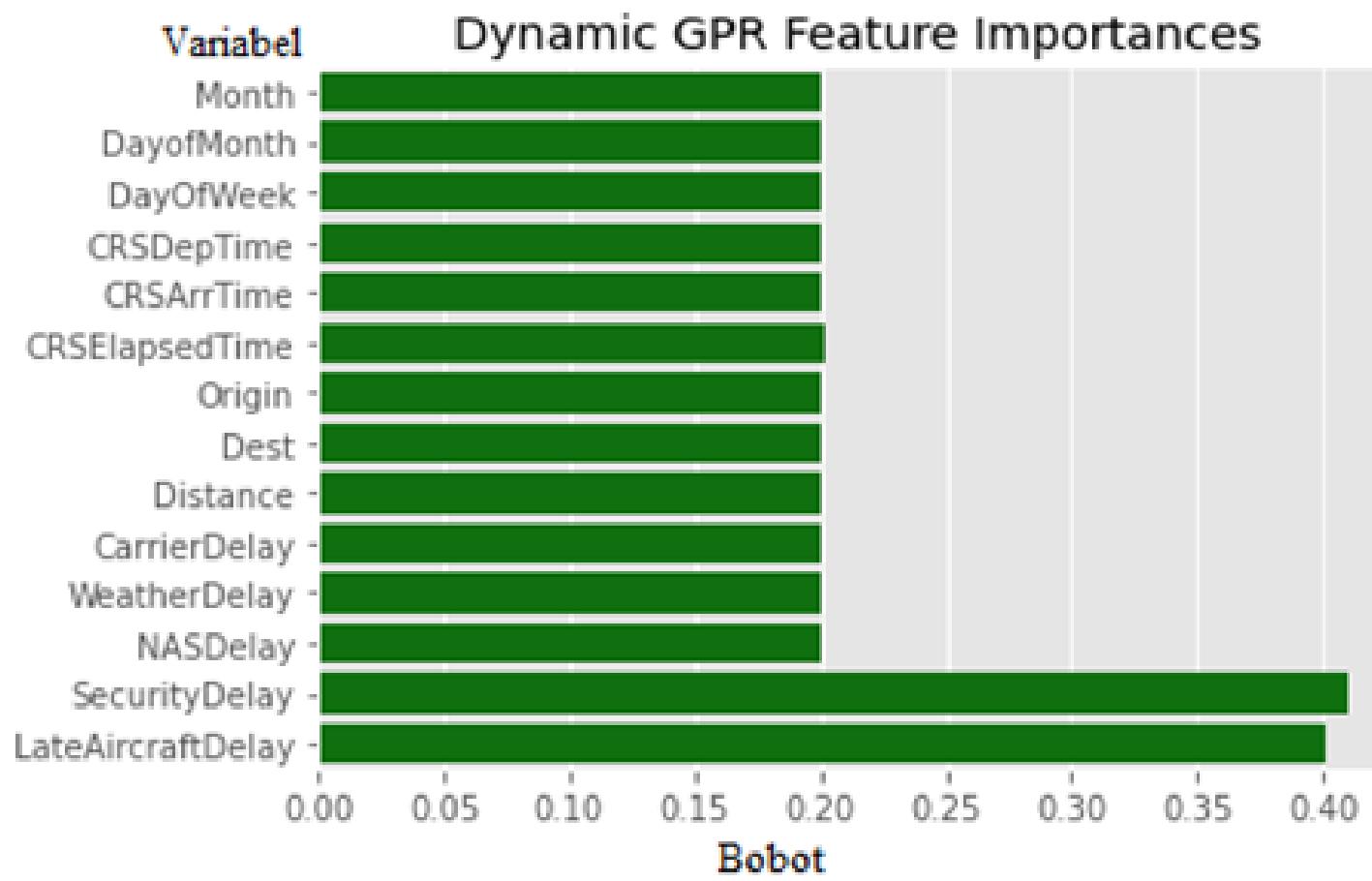
Jumlah Titik Basis 10



Jumlah Titik Basis 25

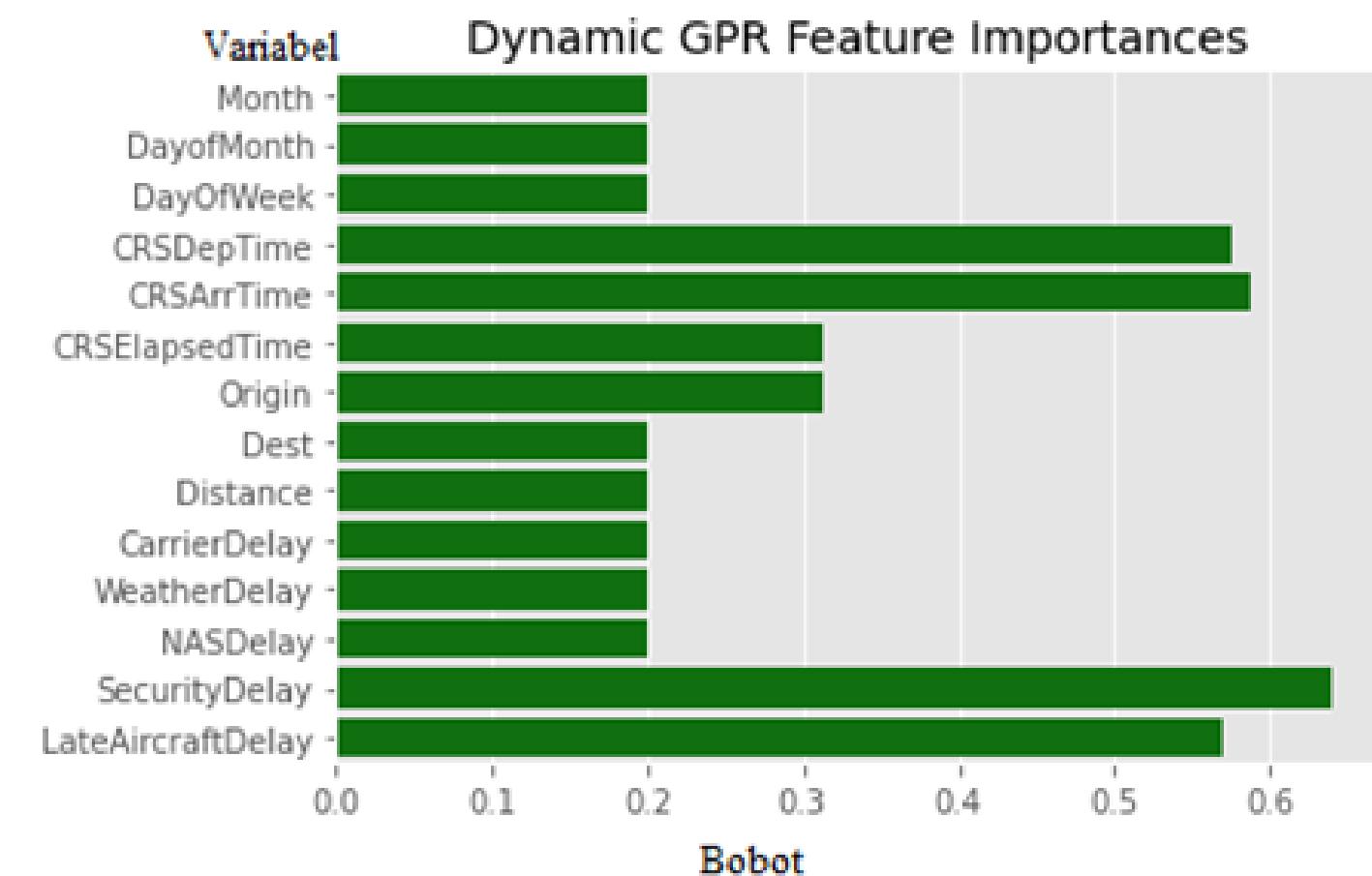
# Hasil Eksperimen

Pada model dengan 10 titik basis, variabel **SecurityDelay** dan **LateAircraftDelay** memegang peran penting dalam memprediksi variabel **DepDelay**.



Jumlah Titik Basis 10

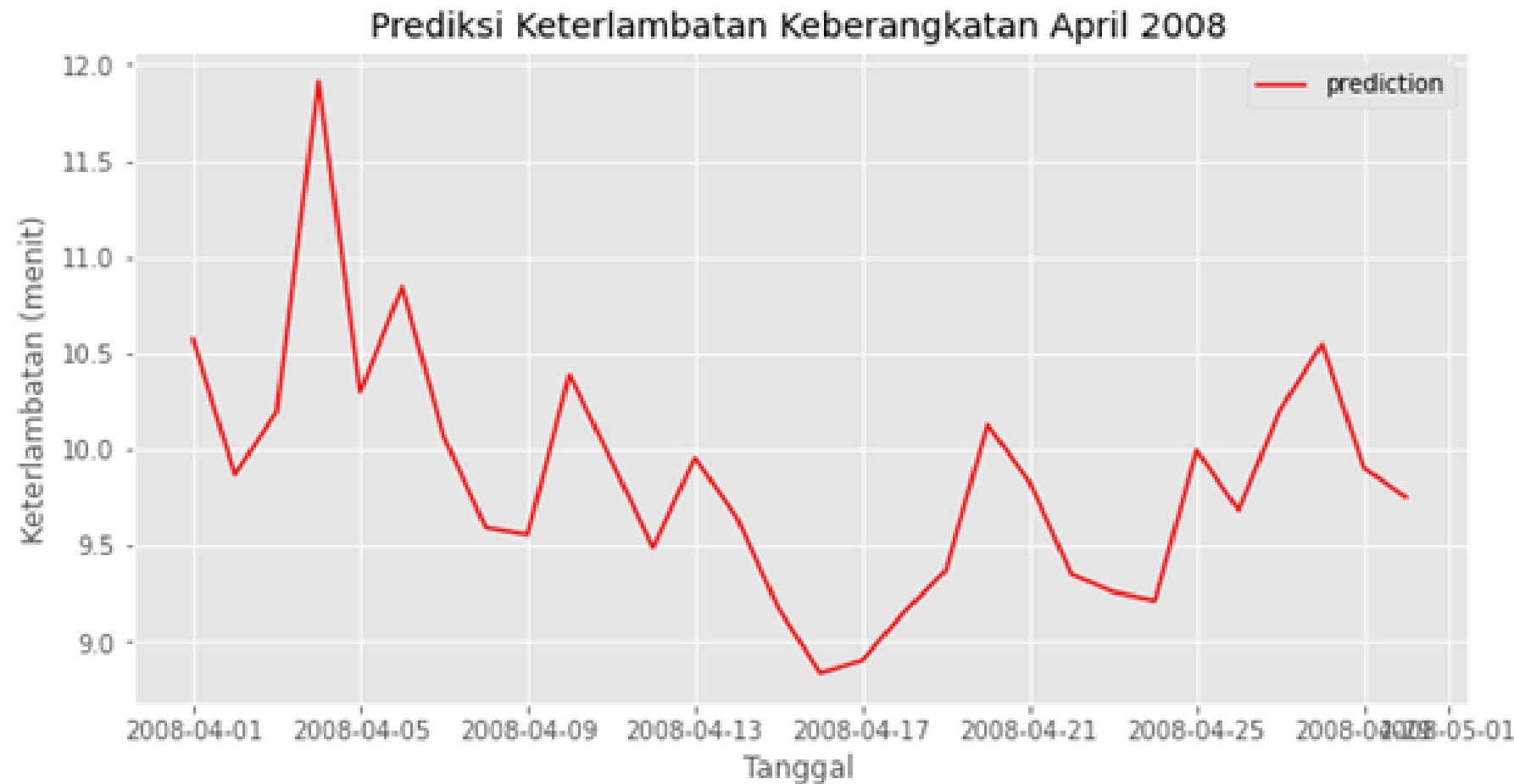
Pada model dengan 25 titik basis, variabel **SecurityDelay**, **CRSArrTime**, **CRSDepTime**, dan **LateAircraftDelay** memegang peran penting dalam memprediksi variabel **DepDelay**.



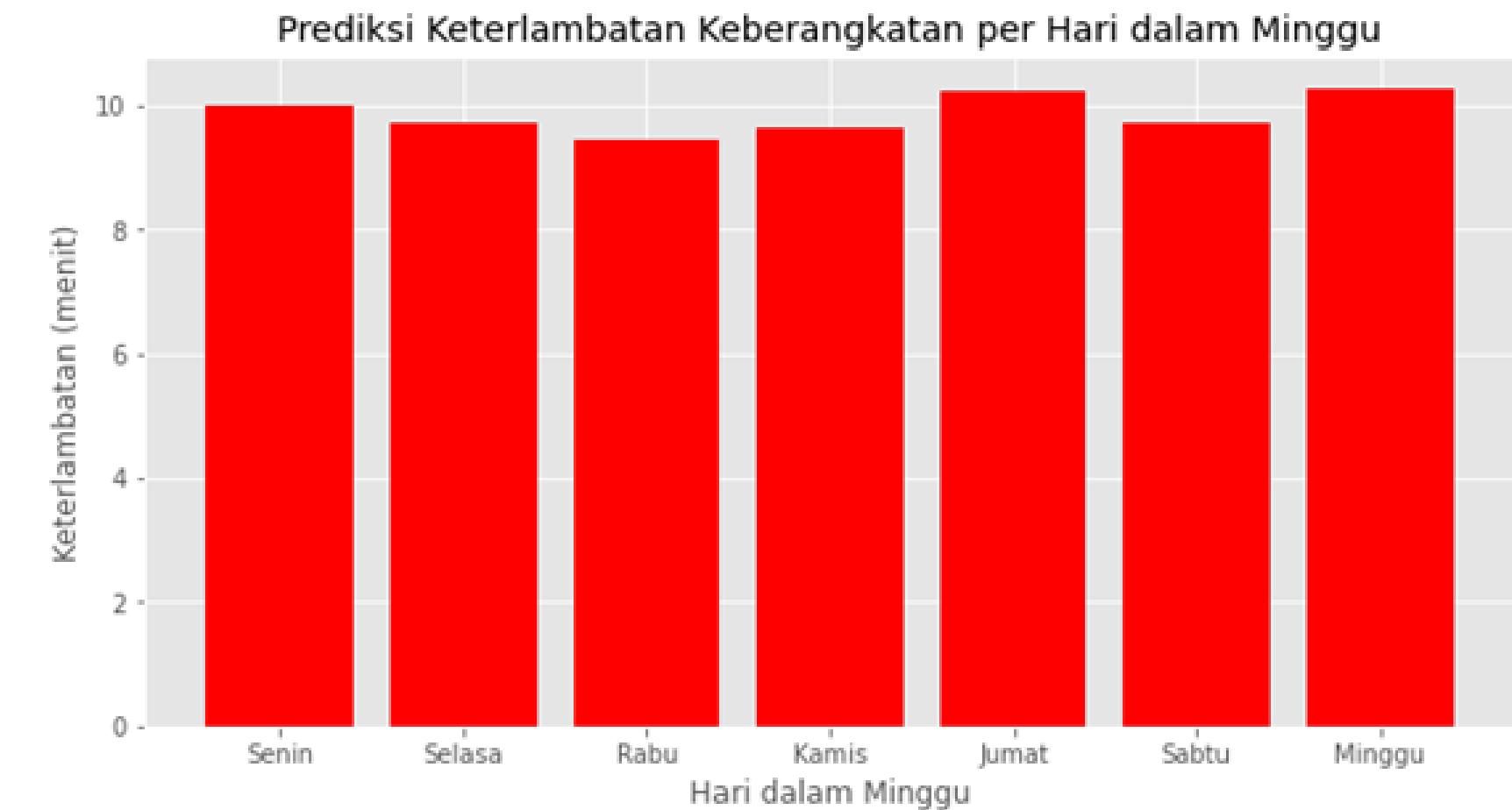
Jumlah Titik Basis 25

# Hasil Eksperimen

Hasil forecast model dengan jumlah titik basis 25 pada bulan April adalah sebagai berikut:



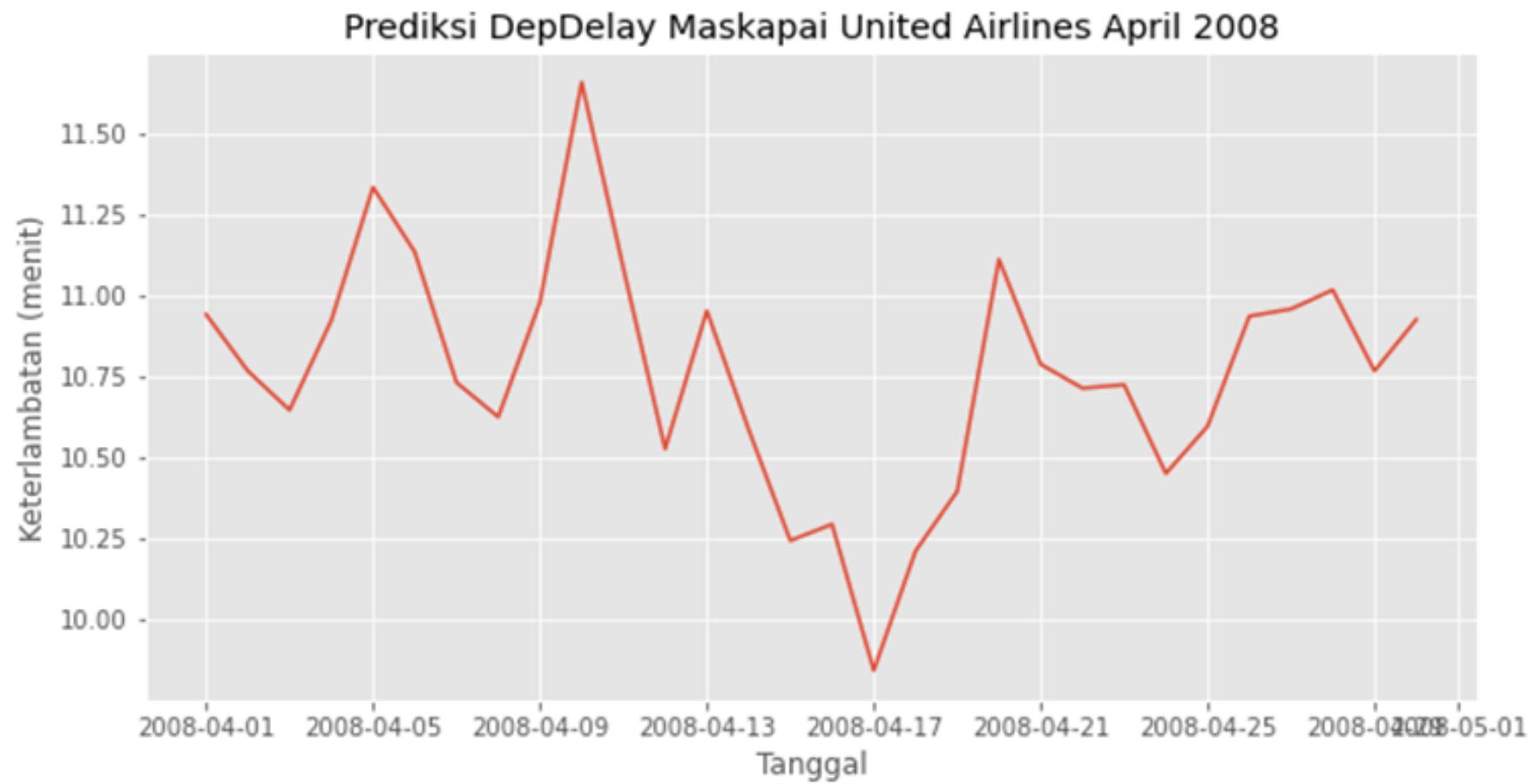
Waktu keterlambatan  
keberangkatan paling rendah  
secara rata-rata akan terjadi pada  
13-17 April 2008



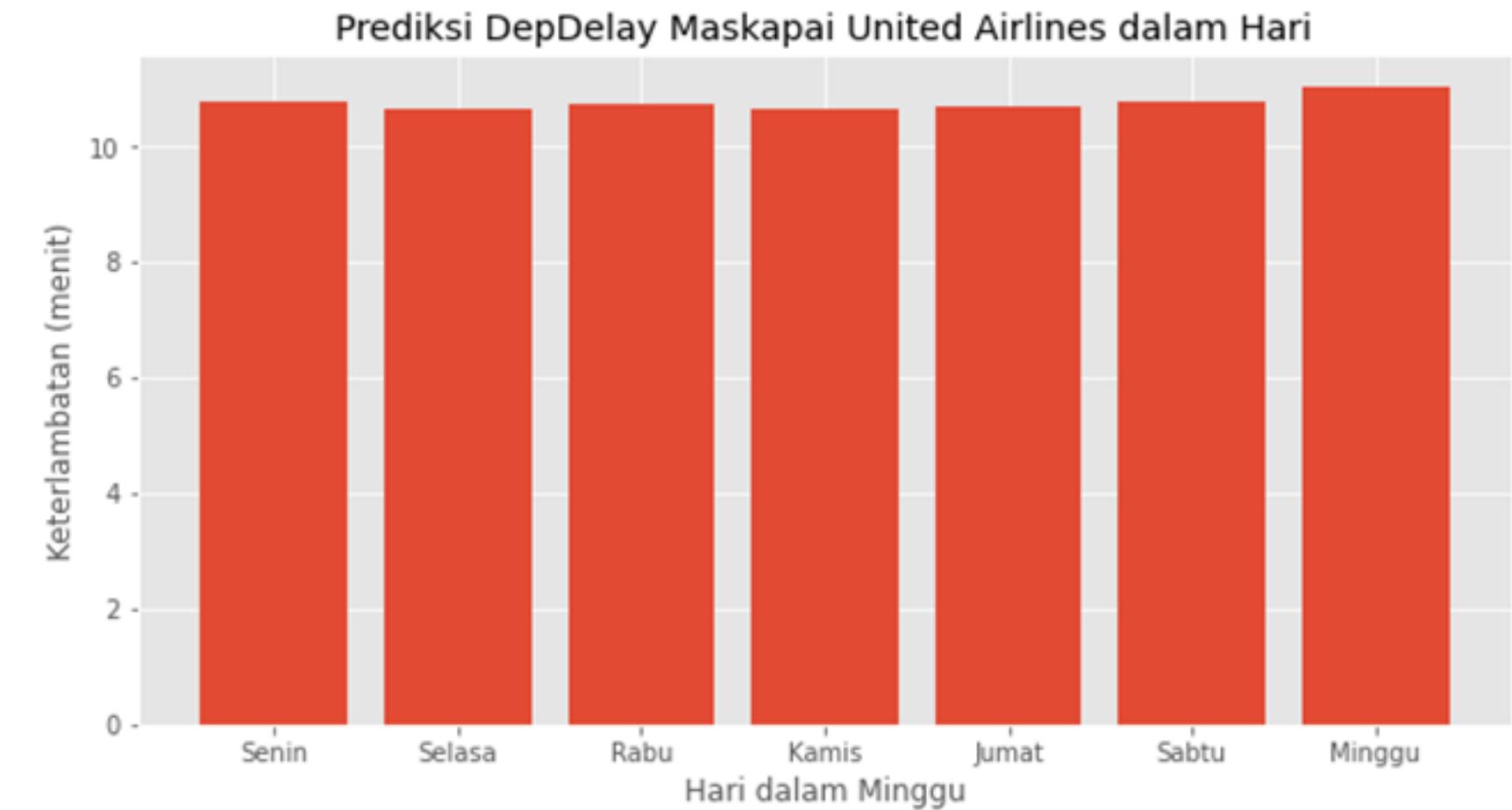
Hari Rabu menjadi hari dengan  
waktu keterlambatan  
keberangkatan paling rendah.

# Hasil Eksperimen

Berikut adalah prediksi keterlambatan keberangkatan untuk maskapai United Airlines



Waktu keterlambatan  
keberangkatan paling rendah  
secara rata-rata akan terjadi pada  
17 April 2008



Hari Selasa, Kamis, dan Jumat  
menjadi hari dengan waktu  
keterlambatan keberangkatan  
paling rendah.

# Kesimpulan

1. Pada bulan April 2008, dari data seluruh maskapai, keterlambatan keberangkatan **paling tinggi** diprediksi terjadi pada tanggal **4 April 2008** dan **paling rendah** terjadi pada **16 April 2008**.
2. Pada bulan April 2008 maskapai United Airlines akan mengalami delay keberangkatan **paling rendah** pada tanggal **17** dan delay keberangkatan **paling tinggi** akan terjadi pada **tanggal 10**.
3. Dari data seluruh maskapai diprediksi hari **Jumat** dan **Minggu** menjadi hari dengan delay keberangkatan **paling tinggi** dan hari **Rabu** merupakan hari dengan delay keberangkatan **paling rendah**.
4. Maskapai United Airlines diprediksi akan mengalami delay keberangkatan **paling rendah** pada hari **Selasa, Kamis dan Jumat**. Sedangkan delay keberangkatan **paling tinggi** terjadi pada hari **Minggu**.
5. Berdasarkan *confusion matrix* diperoleh:

Kategori	Titik Basis 10	Titik Basis 25
<i>True Positive</i>	484.029	484.028
<i>False Positive</i>	0	1
<i>True Negative</i>	2	10.873
<i>False Negative</i>	102.692	91.821



# Referensi

- Beckers, Thomas. (2021). An Introduction to Gaussian Process Models. arXiv preprint arXiv:2102.05497.
- Micci-Barreca, D. (2001). A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. SIGKDD Explor. Newsl., 3(1), 27–32.
- Github Repository:  
<https://github.com/Andikazidanef15/DynamicGPR>



# Terima Kasih

 Search Destination

End Slide

